# Adding Covariate data

# 1. Covariate-dependent hierarchical Dirichlet process

**Authors:** Huizi Zhang, Sara Wade, Natalia Bochkina

**Link:** <https://arxiv.org/abs/2407.02676>

## 1.1 Aim

The paper aims to address the challenges in clustering across related groups when additional covariate information is available. The authors propose a Bayesian nonparametric approach based on mixture models, which combines ideas from the hierarchical Dirichlet process (HDP) and the “single-atoms” dependent Dirichlet process (DDP). The proposed method accommodates both continuous and discrete covariates through appropriate kernel functions, enhancing the model’s flexibility and generality.

C-HDP maintains the same hierarchical structure as HDP. Therefore, it will return the same number of hierarchy layers as HDP.

## 1.2 Data used

### 1.2.1 Type of data

The study primarily uses **single-cell RNA sequencing (scRNA-seq) data**, which is characterized by high-dimensionality. Each data point represents the gene expression levels in a single cell.

### 1.2.2 Data structures

**Gene Expression Matrix:**

The data is structured as a matrix where rows correspond to individual cells and columns to different genes. Each entry in the matrix represents the count of mRNA molecules for a specific gene in a specific cell.

**Covariates:**

Latent Time (tc,d), a continuous variable ranging from 0 to 1. It represents the inferred position of a cell along its developmental trajectory, derived from RNA velocity models. This covariate plays a significant role in the clustering process by influencing the probability of cluster membership.

## 1.3 Methodology:

### 1.3.1 Overall idea

The original HDP is defined as follow:

A close-up of a number

Description automatically generated

A math equations and formulas

Description automatically generated with medium confidence

For each word yj in the document d, we repeatedly sample a θj with probability pj,d. This means that for every single word in the document, they have the same probability of choosing θi, d. However, under C-HDP:

A black symbol with numbers and symbols

Description automatically generated

The probability of drawing θj depends on the word xi, d. pj,d is now constructed as follow:

A math equations and symbols

Description automatically generated with medium confidence

This means that each of the word has its own probability of choosing θj depending on the kernel K which takes in the covariate of each point as input.

**Single atom**: The atoms do not depend on x and are the same as in normal DP. However, the atoms depend on the covariates.

**Single weight**: the weights do not depend on x and are the same as in normal DP. However, the atoms depend on the covariates.

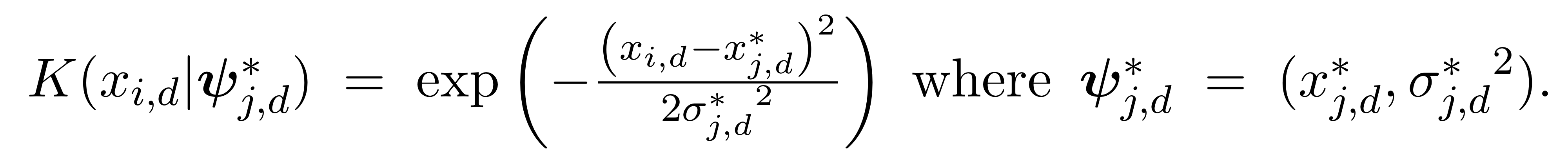
### 1.3.2 Kernel functions

*K* is a kernel function which depends on kernel parameters Ψ and satisfies 0 < *K* < 1.

To make C-HDP to work with various type of covariates, the paper suggests three different kernel function, *K*.

1. **Gaussian Kernel**:

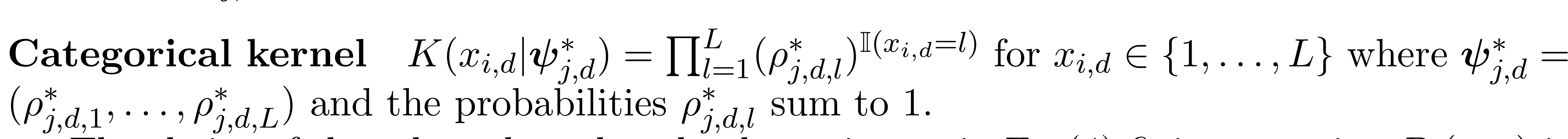
This is used for continuous covariates where the influence on the topic distribution decreases smoothly as covariates become more different.



represents the center and dispersion of the covariate.

**When to Use:** When your covariates are continuous (e.g., time, document length) and you expect a gradual change in topic distribution as the covariates shift.

1. **Categorical Kernel**:



Use this when the covariate is discrete and categorical.

### 1.3.3 Likelihood functions

1. Gaussian likelihood 🡺 For continuous response.
2. Vector autoregressive (VAR) model 🡺 extension of the normal likelihood to time-series data.
3. Negative-binomial likelihood 🡺 For count data.

## 1.4 Similarity

We can incorporate the single-atom DDP part into HDP and the covariate we will be using is the adjacency matrix. However, we will have to look for a suitable kernel function. Else, we can also try to transform the adjacency matrix into something else.

For example:

**Degree Centrality** 🡺 can use negative-binomial likelihood

**Betweenness Centrality**: Measures how often a node lies on the shortest paths between other nodes.

**Closeness Centrality**: Measures how close a node is to all other nodes in the network.

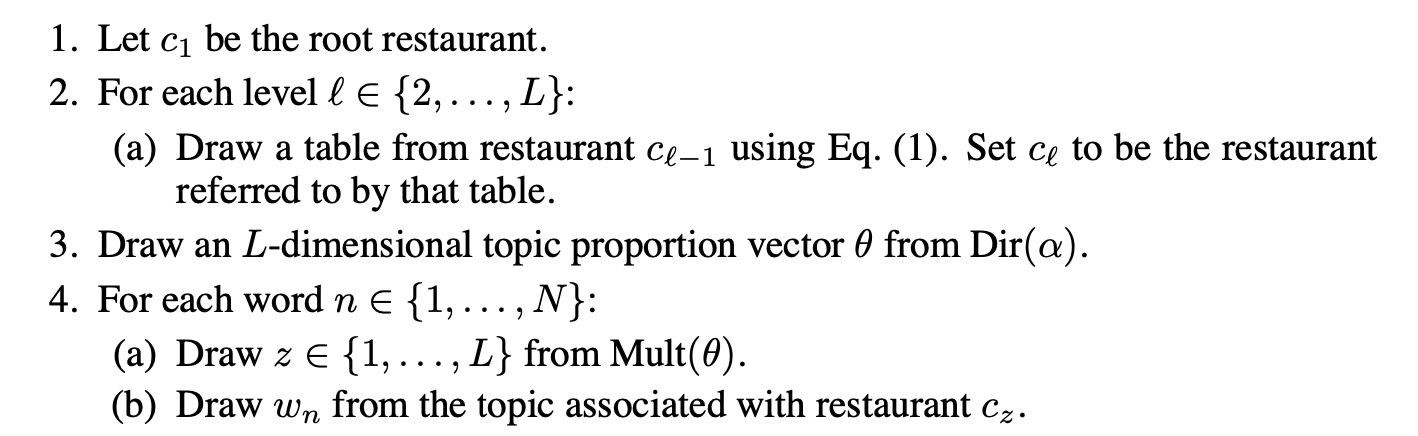
**Eigenvector Centrality**: Measures a node’s influence based on the importance of its neighbours.

# 2. Adding in more layers

## 2.1 Nested Chinese restaurant process(nCRP)

In nCRP, instead of a single-level clustering, the data is organized into multiple nested layers of clusters, where each cluster at one level can generate subclusters at the next level.

The hierarchical structure emerges as customers (data points) move through the hierarchy by choosing tables (clusters) at each level, where each table in a restaurant can open a new restaurant (subcluster) at a deeper level. This process naturally leads to the formation of a tree-like structure.



The generative process is shown above. The only input will be the corpus, and the output will be a **tree-like structure of clusters or topics**, with documents or data points organized into increasingly specific groups as you move deeper into the hierarchy.

## 2.2 Similarity

Exploring the possibility to apply this to each of the groups found such that it gives a hierarchical tree of topics.

Already have some hierarchy layers but still may have some hidden ones. How to apply this algo to such case.

Citation data.